Speech Communication 144 (2022) 57–66

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Recursive Feature Diversity Network for audio super-resolution

Bo Jiang [a], Mixiao Hou [a], Jiahuan Wang [a], Yao Lu [a,*], David Zhang [b,1], Guangming Lu [a,c,**]

[a] Department of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, 518057, China
[b] School of Data Science, Chinese University of Hong Kong, Shenzhen, China
[c] Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies, China

ARTICLE INFO

Keywords:
Audio super-resolution
Lightweight model
Recursive feature diversity
Back-projection block

ABSTRACT

Deep learning methods have been successfully applied to audio super-resolution tasks. Although deep learning methods produce good performance, they are not practical for the real-world applications due to the large member of computations. To address this problem, we propose a Recursive Feature Diversity Networks (RFD-Nets), which is a lightweight model for achieving fast and accurate audio super-resolution. RFD-Nets are composed of a Recursive Feature Diversity (RFD) block and a Back-Projection (BP) block. Specifically, the RFD block is a recursive structure to iteratively refine and extract hierarchical audio feature. Subsequently, using an up-and-down sampling learner, the proposed BP block can effectively capture the deep relationships between High-Resolution (HR) and Low-Resolution (LR) audio pairs, thus producing high-quality audio reconstruction. Furthermore, we collect seven different types of complex audio datasets for training and comprehensively evaluating the proposed method. Extensive experiments demonstrate that our RFD-Nets can achieve superior accuracy on the proposed benchmark datasets against state-of-the-art methods while only requiring lower computation and memory. ***Datasets are released at*** https://github.com/JiangBoCS/RFD-Net.

1. Introduction

Audio super-resolution aims to recover the corresponding high-quality High-Resolution (HR) audios from the Low-Resolution (LR) audios. Audio super-resolution is evolved from the task of improving the listening quality of narrowband speech (Epps and Holmes, 1999). Recently, audio super-resolution can be elegantly embedded into many real-world applications such as speech recognition and speech synthesis (Li et al., 2015). In the future, the mature audio super-resolution model is expected to replace the recording facility of lossless music, thus greatly saving resource costs.

In the past few years, audio super-resolution has been also known as bandwidth expansion and reconstructed signals based on the frequency domain. For example, Park and Kim (2000) transformed narrowband spectral envelope to wideband spectral envelope based on linear predictive coding with Gaussian mixture. Jax and Vary (2003) estimated the spectral envelope of the extension hand based on a Hidden Markov Model.

Recently, with the great achievements of deep learning in the field of image super-resolution (Agustsson and Timofte, 2017; Guo et al., 2020; Zhang et al., 2020; Li et al., 2019b; Mei et al., 2020), neural networks have been gradually used to deal with the difficult audio super-resolution task. Deep learning has paved the way for audio super-resolution in both frequency and time domains. Li et al. (2015) proposed a Deep Neural Network (DNN) based bandwidth expansion system to estimate the entire wideband spectrum. Kuleshov et al. (2017) firstly implemented convolutional architectures and introduced a deep residual network to reconstruct high-quality audio on raw signals. Lim et al. (2018) considered the time and frequency domain and developed Time–Frequency Network (TFNet) to solve the problem of recovering low-quality audio. Inspired by Generative Adversarial Networks (GANs) achieving impressive performance in image processing (Zou et al., 2020; Lata et al., 2019; Tang et al., 2020), Eskimez and Koishida (2019) leveraged the GANs for audio super-resolution by generating a log power spectrogram.

Generally speaking, frequency-domain and time-domain-based models both can transform audio from LR to HR. Nevertheless, for the frequency-domain or the time-domain approaches, raw audio signals are always processed using spectral or hand-crafted features for super-resolution. However, in practical scenarios, these methods are not flexible in practical application, and the generalization performance of

* Corresponding author.
** Corresponding author at: Department of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, 518057, China.
E-mail addresses: jiangbo_PhD@outlook.com (B. Jiang), mixiaohou@163.com (M. Hou), 465909954@qq.com (J. Wang), luyao2021@hit.edu.cn (Y. Lu), davidzhang@cuhk.edu.cn (D. Zhang), luguangm@hit.edu.cn (G. Lu).
[1] David Zhang is with the School of Science and Engineering, The Chinese University of Hong Kong at Shenzhen, Shenzhen 518172, China.

https://doi.org/10.1016/j.specom.2022.08.005
Received 12 April 2022; Received in revised form 9 August 2022; Accepted 20 August 2022
Available online 30 August 2022

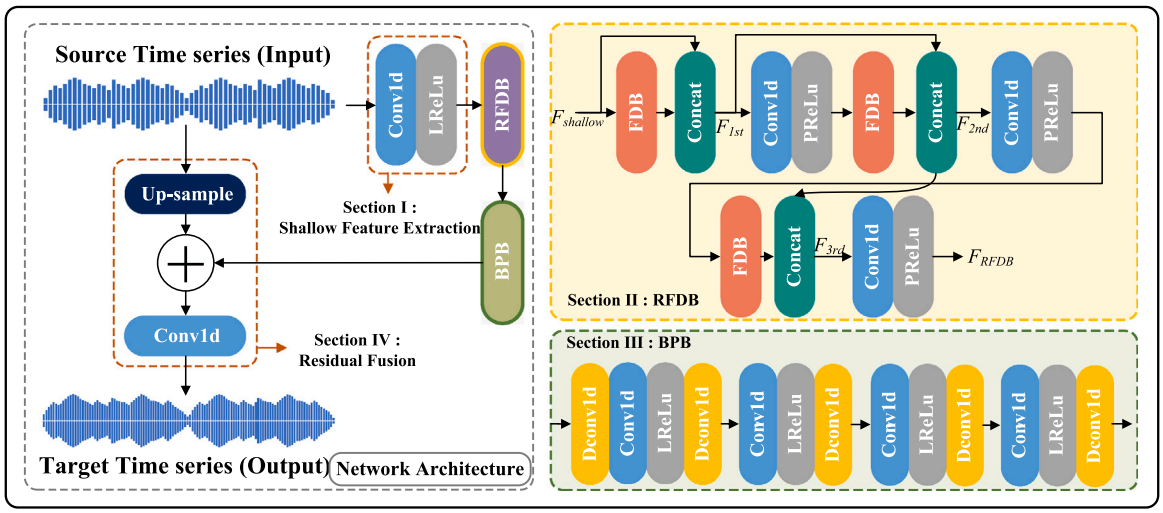

Fig. 1. Overall framework of Recursive Feature Diversity Networks (RFD-Nets) for audio super-resolution. Our model mainly contains two blocks, namely the Recursive Feature Diversity (RFD) block and the Back-Projection (BP) block. The main function of RFD block is to extract more types of feature maps with as few parameters as possible. BP block can accurately capture the mutual mapping relationship between LR-HR audio pairs.

these methods is limited. Although the image denoising performance of DNN-based methods has been significantly improved compared to traditional methods, they usually suffer from a critical problem in that most models have huge computation and memory consumption. Therefore, we tend to design an end-to-end efficient audio super-resolution model suitable for applications in real-world scenarios. From this perspective, it is very important to design lightweight deep learning models suitable for practical applications. Additionally, in terms of evaluating the performance and robustness of the model, the current datasets for audio super-resolution are limited and most of them are pure human voices, which is hard to distinguish the pros and cons of the model.

To address the above issues, we propose fast and lightweight Recursive Feature Diversity Networks (RFD-Nets) for audio super-resolution and build a diversity dataset to sufficiently evaluate ASR models. RFD-Nets leverage recursive learning manner to realize effective reconstruction from LR to HR signals. RFD-Nets consist of two recursive blocks, one is the Recursive Feature Diversity (RFD) block and another is Back-Projection (BP) block. Particularly, the core component of the RFD block is the Feature Diversity (FD) block with a cost-efficient structure, which can refine and extract multiple types of features. FD block is a dichotomous structure that mainly obtains more types of feature maps for improving the audio quality of reconstruction without changing the number of channels. Noticeably, we use a recursive operation on the FD block (*i.e.*, weight sharing) in the RFD block. Benefiting from these recursive modules, the number of parameters is greatly reduced to time up the inference process of our model. Moreover, the BP block is a module inspired by Larsen and Aarts (2005), which is an efficient up-and-down sampling manner in signal reconstruction. Integrating the proposed RFD block with the BP block, our RFD-Nets can capture the deep relationships between HR and LR pairs and thus realize higher-quality audio reconstruction results. In this work, the newly collected audio super-resolution dataset provides more types of samples, such as music fragments of a multi-instrument ensemble, *etc.*, which can effectively evaluate the performances of various models. The main contributions of this paper can be summarized as follows:

- This work proposes a fast and lightweight Recursive Feature Diversity Networks (RFD-Nets) for audio super-resolution by introducing Recursive Feature Diversity (RFD) block and Back-Projection (BP) block for maximizing the diversity learning of features and maintaining the quality of audios in accuracy.
- We construct the RFD block with recursion structure and the BP block with the up-and-down sampling structure to extract the deep relationships between HR and LR audio pairs, which steadily and significantly improves the audio super-resolution performance.
- We build a new complicated dataset for audio super-resolution, which covers multi-instrument playing, multiple vocals and dual-track music. Extensive experiments show that the proposed model clearly outperforms state-of-the-art baselines with a larger learning rate.

The paper is organized as follows: Section 2 introduces the related works. Section 3 presents the proposed RFD-Nets. Section 4 demonstrates the experiment results, and the paper is finally concluded in Section 5.

2. Related work

Audio super-resolution, aims to recover high-resolution signals (especially high-frequency details and content) from low-resolution audio signals. At the same time this task is also known as bandwidth scaling (Larsen and Aarts, 2005; Ekstrand, 2002). Early traditional methods estimate high-resolution signal spectral parameters from low-resolution audio signals, such as Hidden Markov Models (Jax and Vary, 2003), Linear Predictive Coding (Bachhav et al., 2018), Nonnegative Matrix Factorization (Bansal et al., 2005), and Gaussian Mixture Models (Seo et al., 2014). Compared with traditional methods, convolutional neural networks can overcome the inflexible limitations of manual design, thereby enhancing the representation ability of audio signals (Jiang et al., 2022b,a,c).

The first framework to employ convolutional neural networks for audio super-resolution reconstruction was proposed by Li and Lee (2015). Inspired by the task of image super-resolution reconstruction, Kuleshov et al. (2017) proposed to reconstruct high-resolution audio signals based on convolutional encoder–decoder networks. WaveNet (Shen et al., 2018) uses the dilated convolutions to preserve the original resolution while increasing the receptive field in the network to improve reconstruction performance. FFTNet (Feng et al., 2019) uses a frequency domain transformation process with the classical fast Fourier transform in a convolutional neural network to reconstruct high-resolution audio signal details. Although existing deep neural networks have achieved remarkable breakthroughs in audio super-resolution, the computational burden is too heavy, which greatly hinders the application in practice.

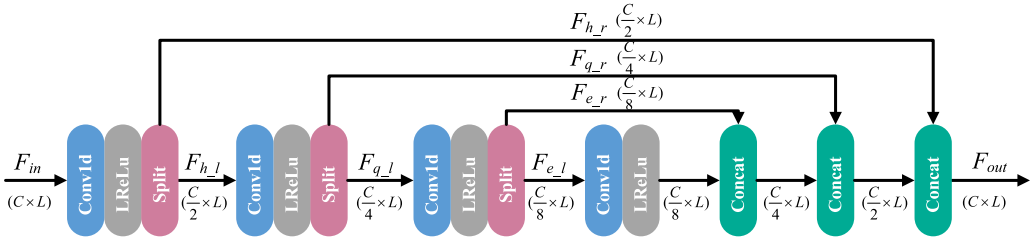

Fig. 2. FD block: Feature Diversity Block. The split operation divides the number of channels of feature maps into two equal parts, and delivers them to different modules, respectively.

Different from the deep learning methods mentioned above, we propose a Recursive Feature Diversity Network (RFD-Nets) including Recursive Feature Diversity (RFD) block and Back Projection (BP) block, which is flexible and efficient for lightweight audio super-resolution tasks. RFD block is a recursive structure that can refine and extract many types of features. Subsequently, combining with the up-and-down sampling structure of BP block, the model can effectively capture the deep relationship between high-resolution (HR) and low-resolution (LR) audio pairs, resulting in high-quality audio reconstructions.

3. Proposed method

3.1. Problem definitions

The audio super-resolution task is the process of recovering high-resolution (HR) audio with a sample rate R_2 from a given low-resolution (LR) audio with a sample rate R_1. We use LR to represent low-resolution audios and HR to represent high-resolution audios. We use Eq. (1) to express the relationship between HR audios and LR audios:

$$S_{LR} = \mathcal{H}\left(S_{HR};\eta\right), \tag{1}$$

where $\mathcal{H}$ denotes a degradation mapping function, S_{HR} represents HR audio and η is the parameters of the degradation process. We assume that this degradation model is directly obtained by the downsampling operation, as follows:

$$\mathcal{H}\left(S_{HR};\eta\right) = \left(S_{HR}\right)\downarrow_s, \{s\} \subset \eta, \tag{2}$$

where $\downarrow_s$ is a downsampling operation with the scaling factor $s = \frac{R2}{R1}$. ***Therefore, audio super-resolution reconstruction is an ill-posed problem.*** At this point, we need to get an approximation $\hat{S}_{HR}$ (HR approximation) of the true high-resolution audio:

$$\hat{S}_{HR} = f\left(S_{LR};\theta\right), \tag{3}$$

where f represents the recovery model of super-resolution audio, and θ is the parameters of the f function. To this end, the optimization objective of audio super-resolution is formulated as follows:

$$\hat{\theta} = \arg\min_{\theta} \mathcal{L}(\hat{S}_{HR}, S_{HR};\theta), \tag{4}$$

where $\mathcal{L}(\hat{S}_{HR}, S_{HR})$ denotes the loss function between the reconstructed HR audio $\hat{S}_{HR}$ and the corresponding ground truth audio S_{HR}. θ is the parameters of degradation model. The loss function for audio super-resolution is element-wise mean absolute error (*i.e.*, time series loss) in this work, which will be illustrated in Section 3.3.

3.2. Network architecture

In this section, we primarily introduce the overall structure and various components of the proposed RFD-Nets, which is shown in Fig. 1. The framework is composed of four parts: (1) shallow feature extraction, (2) Recursive Feature Diversity (RFD) block, (3) Back-Projection (BP) block, and (4) residual fusion.

Low-resolution audio is fed into the shallow feature extraction to extract rich shallow features. Then, the shallow features are transferred to the RFD block, and more types of feature maps are obtained without changing the number of feature channels. Due to using a recursive operation (*i.e.*, the FD block is reused) in RFD block, the efficiency of feature extraction can be significantly enhanced. Subsequently, a large number of features are sent to the BP block. Because the BP block is an up-down-up sampling structure, it can effectively detect the mutual mapping relationship between LR-HR audio pairs. Finally, the learned residual information and the up-sample LR audio information are delivered through the residual fusion part to obtain the final reconstructed audio.

For the given S_{LR} audio, the shallow feature extraction step is given as:

$$F_{shallow} = \sigma\left(H_{shallow}\left(S_{LR}\right)\right), \tag{5}$$

where $H_{shallow}$, σ and $F_{shallow}$ represent the 3×1 convolution, the LeakyReLU activation function and resulting output, respectively.

The structure of RFD block is shown in Fig. 1. For discovering more abundant or deeper feature extraction step, we apply the structure of RFD block as follows:

$$F^{1st} = T_{concat}\left(H_{FDB}\left(F_{shallow}\right), F_{shallow}\right), \tag{6}$$

$$F^{2nd} = T_{concat}\left(H_{FDB}\left(\sigma\left(H_{convA}\left(F^{1st}\right)\right)\right), F^{1st}\right), \tag{7}$$

$$F^{3rd} = T_{concat}\left(H_{FDB}\left(\sigma\left(H_{convB}\left(F^{2nd}\right)\right)\right), F^{2nd}\right), \tag{8}$$

$$F_{RFD} = \sigma\left(H_{convC}\left(F^{3rd}\right)\right), \tag{9}$$

where H_{FDB} represents the mapping function of the FD block. In order to reduce the number of network's parameters, we reuse the FD block (*i.e.*, sharing parameters for multiple times). T_{concat} is the concatenation operation. H_{convA}, H_{convB} and H_{convC} respectively represent different 3×1 convolution operations. F_{RFD} is the feature maps finally generated after RFD block.

Feature Diversity (FD) block: The RFD block is composed of FD blocks and recursive operations. The RFD aims to extract feature maps that are more conducive to reconstruction with a small amount of model parameters. The detailed structure of the FD block is shown in Fig. 2. Specifically, the input feature map is first refined using a convolution layer and an activation function, and then the refined features are divided into two equal parts along the channel axis using the Split Layer. One part is kept, and the other part is sent to the next stage of convolutional layers and activation functions. Finally, all the kept and processed feature maps are concatenated along the channel axis together. The RFD block can be formed by recursing the above FD block multiple times. The RFD block can effectively generate hierarchically rich features with a constant number of channels, making it easier to extract features that are more conducive to reconstruction with a small model parameters. Thus, RFD is more efficient for audio super-resolution reconstruction. *Meanwhile, compared with the traditional recursive program, the proposed RFD block does not*

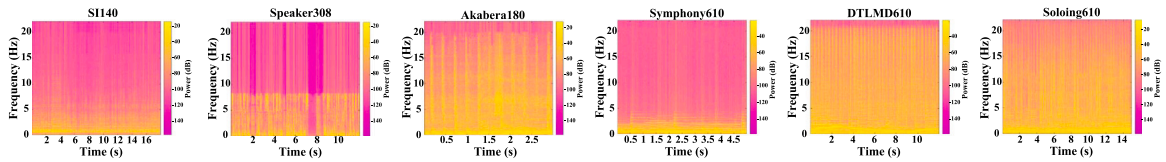

Fig. 3. Visualization spectrum of test dataset example. Since there are only subtle differences between the visualized spectrograms of double track audio from DTLMD610, we only show the spectrogram of one track. The amount of information for audios from SI140 and Symphony610 is mainly concentrated in the low frequency, and the details of high frequency are less. Compared with the above two datasets, there is more information of audios from Speaker308 centralized low frequency. Details of audios from DTLMD610, Akabera180 and soloing610 are distributed in the high frequency part.

need to perform additional scheduling overhead. Our RFD block is a network structure, which can be regarded as a simple sequential shared FD block. Additionally, the FD module, the core component in the RFD, has a small number of parameters. Also, there is a limit number of recursive operations (i.e., shared parameters) used in the RFD module. Therefore, the overall model parameters are few and the computational complexity is limited.

Formally, given an input feature maps $\boldsymbol{F} \in \mathbb{R}^{B\times C\times H\times W}$ obtained by the shallow feature extraction part, the specific above process is written as follows:

$$F_{h\_r},\ F_{h\_l} = \Im\left(\sigma\left(H_r\left(F_{in}\right)\right)\right), \tag{10}$$

$$F_{q\_r},\ F_{q\_l} = \Im\left(\sigma\left(H_q\left(F_{h\_r}\right)\right)\right), \tag{11}$$

$$F_{e\_r},\ F_{e\_l} = \Im\left(\sigma\left(H_e\left(F_{q\_r}\right)\right)\right), \tag{12}$$

$$F_{last} = \sigma\left(H_{last}\left(F_{e\_r}\right)\right), \tag{13}$$

$$F_{out} = T_{concat}\left(F_{h\_l},\ F_{q\_l},\ F_{e\_l},\ F_{last}\right), \tag{14}$$

where F_{in} is the input feature maps, F_{h_l}, F_{h_r}, F_{q_l}, F_{q_r}, F_{e_l}, F_{e_r} respectively denote different refined feature maps. H_r, H_q, H_e and H_{last} are different 3×1 convolution operations, respectively. $\Im$ represents the split operations for the channel of feature maps.

Back-Projection (BP) block: Back-projection (Irani and Peleg, 1991) minimizes reconstruction errors, which was widely applied to image processing (Haris et al., 2017; Zhao et al., 2017). The back-projection, however, has not been applied to the audio field (Haris et al., 2017; Zhao et al., 2017; Dong et al., 2009). Therefore, we extend this back-projection to audio super-resolution tasks. The proposed BP block with an up-down-up sampling structure, is shown in Fig. 1. The BP tries to iteratively apply back-projection refinement, which aims to compute the reconstruction error and fuse it back to tune the HR audio intensity. Specifically, HR and LR audios are related by an iterative up-and-down-projection unit, in which the up-projection unit generates HR features, then the down-projection unit projects the HR features back into the LR spaces. By learning up-sampling (*i.e.*, implemented with deconvolutional layers) and down-sampling (*i.e.*, implemented with convolutional layers) modules, this schema allows the networks to preserve the HR components and construct abundant LR features and HR features. Therefore, such a structure can adjust the reconstruction error between HR-LR audio, improving the quality of reconstructed audio.

The proposed BP block tries to iteratively apply back-projection refinement, which aims to improve the quality of reconstructed audio. The proposed BP block with an up-down-up sampling structure can effectively generate the features with different hierarchy, so that the prior features required for audio reconstruction can be produced more easily. Therefore, the proposed BP block can generate rich reconstruction features under the condition of a limited amount of parameters, to improve the performance of audio reconstruction tasks with a lightweight model in advance.

The RFD block can not only reduce its parameters through recursive operations, but also provide rich features for the lightweight model of speech reconstruction. In addition, the proposed BP block can effectively generate layer-rich features with the same number of channels, generating effective prior features for audio reconstruction. ***The above RFD block and BP block are integrated to form the proposed RFD-Nets, which can not only reduce the number of model parameters, but also ensure the performance of speech reconstruction.***

Residual Fusion: Since the audio super-resolution is an audio-to-audio nonlinear regression task, where the input audio is highly correlated with the target audio, we try to learn the residuals between them by using deep 1-D convolutional neural network. In this case, it avoids learning a complicated transformation from complete audio to another, instead of only learning the residual information to restore the missing high-frequency details. The residual fusion process is shown in section IV of Fig. 1. When reconstructing high-resolution audio, we implement LR audio upsampling to supplement the low-frequency information in the reconstruction process for complementing the missing low-frequency information. The final reconstructed high-resolution audio is obtained by adding the upsampled information from LR audio and the residual information learned of the network.

3.3. Loss function

Mean Absolute Error (MAE) loss is the most frequently used loss functions for nonlinear regression tasks (including the audio restoration task). In this work, we adopt the MAE loss for measuring the differences between the SR audios and the ground truth. Specifically, the loss function is defined as follows:

$$Loss = \frac{1}{M}\sum_{i=1}^{M} \| S_{GT}^{i} - \mathcal{F}\left(S_{LR}^{i}\right) \|_1, \tag{15}$$

where S_{GT}^{i} and S_{LR}^{i} denote the ith LR audio and the corresponding ground truth, respectively M is the total number of training samples. $\mathcal{F}(\cdot)$ represents the RFD-Nets.

4. Experiments

4.1. Implementation details

Our lightweight model RFD-Nets consists of the RFD block module and the BP block module. We recursively use the FD block module three times in RFD block, reducing the number of parameters by about twice. In BP block, the up-down-up structure is also used three times. The upsample and downsample rates both are $s = \frac{R2}{R1}$ in BP block. In RFD-Nets, the kernel size of all convolutional layers is 3×1, except for the last convolutional layer, which has a kernel size of 1×1. The activation functions are LeakyReLU. We used Adam optimizer (Kingma and Ba, 2014) to train our models with initial learning rate 1×10^{-4}. The models are trained for 300 epochs. The learning rate decays every 60 epochs, and the decay factor is 0.1. We set the batch size to 64 and implement our network on PyTorch (Paszke et al., 2017). We train it on NVIDIA RTX TITAN GPU.

Table 1
Basic information about the collected datasets, including the number of segments in each dataset, the average time per segment, and the type of each dataset, such as training dataset or test dataset.

Dtatsets	Segment	Time	Application type
CCMD	1500	1 s	Train
DTLMD610	610	11 s	Test
Symphony610	610	5 s	Test
Solosing610	610	15 s	Test
Akabera180	180	3 s	Test
SI140	140	18 s	Test
Speaker308	308	12 s	Test

4.2. Dataset

Currently, there are very few audio super-resolution datasets available. In previous work, VCTK dataset (Eskimez and Koishida, 2019; Kuleshov et al., 2017) has been frequently used to evaluate the audio super-resolution models, but this dataset only contains the voices of people talking in single styles. Therefore, we collected and established the following diverse datasets for the audio super-resolution task:

CCMD. For training the proposed network, we build the Chinese Classical Music Dataset (CCMD), which consists of ancient music played by Chinese unique musical instruments. This dataset has a total of 1500 LR/HR audio pairs.

To test the generalization performance of the model, we also provide six datasets with different attributes. The tested dataset details are as follows:

DTLMD610. Double track light music dataset (DTLMD) consists of 610 high-quality LR/HR audio pairs. Audio in this dataset has two tracks and various styles of light music (without vocals). This is the first attempt to evaluate the audio super-resolution model with the dual-track music dataset.

Symphony610. This dataset contains 610 symphony fragments, including concertos, symphonic poems, and symphonic overtures. It has the characteristics of complex orchestral instruments.

Solosing610. This dataset contains 610 LR/HR audio pairs. Audios in this dataset are acted by different singers. The complex and varied styles of the voices and accompaniments of different singers will pose a great challenge for the audio super-resolution task.

Akabera180. The audios of this dataset are sung by the pure human voice, which means the accompaniment is also imitated by the human voice. This dataset covers 180 LR/HR audio pairs.

SI140. This dataset contains 140 LR/HR audio pairs, which is composed of solo pieces from different instruments without any accompaniment.

Speakers308. This dataset has 308 LR/HR audio pairs, and it includes speeches of people at different ages.

These datasets are listed in Table 1 in detail and we also visualized their spectrograms for observing the characteristics of them in Fig. 3. Since there are only subtle differences between the visualized spectrograms of double track audio from DTLMD610, we only show the spectrogram of one track. From Fig. 3, we can clearly figure out the differences between these datasets. The amount of information for audios from SI140 and Symphony610 is mainly concentrated in the low frequency, and the details of high frequency are poor, so these datasets are easier to restore. Compared with the above two datasets, there is more information of audios from Speaker308 centralized low frequency. Details of audios from DTLMD610, Akabera180 and soloing610 are distributed in the high frequency part, which will bring more challenges for audio super-resolution.

The sampling rate of HR audios in the above datasets is 44,100 Hz. The LR audios are obtained from the corresponding HR audios by an interpolate function with default settings in Pytorch (*i.e.*, linear interpolation). Notably, for the DTLMD610 dataset, we regard dual-track audio as a combination of two independent single-track audio, so when processing dual-track audio, we put their single-track audio into the audio super-resolution model for super-resolution reconstruction. We recombine the reconstructed single-track audio to obtain reconstructed dual-track audio.

4.3. Objective metrics

We adopt three metrics to measure the effect of the model: Peak Signal-to-Noise Ratio (PSNR), Signal-to-Noise Ratio (SNR) and Log-Spectral Distance (LSD) (Gray and Markel, 1976). Among them, PSNR and SNR are commonly used metric. The higher value of PSNR or SNR implies higher quality of the generated audio. The SNR measures the reconstruction quality of individual frequencies as Eq. (16).

$$\mathrm{SNR}(S_{SR}, S_{HR}) = 10\log_{10}\frac{\|S_{HR}\|_2^2}{\|S_{HR} - S_{SR}\|_2^2}, \tag{16}$$

where S_{SR} and S_{HR} are the estimated log-power and ground truth spectrogram, respectively.

The specific calculation of LSD is shown in Eq. (17).

$$\mathrm{LSD} = \frac{1}{L}\sum_{l=1}^{L}\sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(S_{HR}(l,k) - S_{HR}(l,k)\right)^2}, \tag{17}$$

We use l and k index frames and frequencies ($K = 2048$).

4.4. Experiment comparisons

The proposed method is evaluated in comparison with state-of-the-art algorithms. To be fair in comparison, we produce their results using public-available implementations provided by the reported settings in the corresponding literatures. Meanwhile, to demonstrate that the proposed RFD-Nets has stronger robustness and generalization performance, we also conduct a comparative evaluation with the state-of-the-art algorithms in the field of audio super-resolution and bandwidth extension.

Comparison with baseline methods In the comparisons of the audio super-resolution task, we compare the proposed RFD-Nets with baseline methods (*i.e.*, Spline Schoenberg, 1973, DNN Li et al., 2015 and AudioUNet Kuleshov et al., 2017) on the six benchmark datasets collected in this work. Table 2 presents the average results of PSNR, SNR and LSD on six benchmarks datasets with different scale factors (*i.e.*, scale $= 2, 4, 6$). From this table, it can be clearly seen that these compared methods generally perform the best on datasets of instrumental music (SI140 and Symphony610), followed on the datasets of single speaker and dual-track music (Speaker308 and DTLMD610), and finally on the datasets of complex vocals with music (Akabera180 and Solosing610). As the complexity of the audio datasets changes, our model still maintains superiority in restoration tasks at different scales. It can be observed that our model has a greater advantage than the comparison models in processing dual-track audio. Compared with DNN, the average PSNR of our RFD-Nets on the six benchmark datasets not only improves by about 7.53 dB, but also reduces the amount of parameters by about **52.23 times**. Moreover, our RFD-Nets achieves the best results compared to the AudioUNet, while the model size of RFD-Nets are reduced about **15.38 times** than AudioUNet. This indicates our RFD-Nets can improve the learning ability of the audio super-resolution networks, leading to significantly reducing the parameters. Figs. 4, 5, and 6 shows the spectrogram comparisons of different methods on six benchmark datasets with scale $= 2, 4, 6$, respectively. The compared methods can not only reconstruct the missing frequencies but also destroy the spectrogram details, resulting in over-smoothness, artifacts and severely smearing in many energy areas. However, the visualization result achieved by the proposed RFD-Nets is the closest to that of ground truth without introducing other artifacts. These results fully show that the proposed RFD-Nets achieve a strong generalization for audio super-resolution spectrograms.

Table 2
The comparison results for ×2, ×4 and ×6 audio super-resolution experiments. The values of PSNR and SNR are positively correlated with visual quality, while the value of LSD is negatively correlated with visual quality.

Method	Scale	Params	DTLMD610			Symphony610			Solosing610		
			PSNR	SNR	LSD	PSNR	SNR	LSD	PSNR	SNR	LSD
Spline		–	39.83	37.13	2.61	49.94	39.91	1.21	34.46	33.16	3.63
DNN		148.1M	40.65	38.46	2.52	50.35	41.36	1.19	35.61	35.98	3.52
AudioUNet	×2	36.9M	41.85	40.22	2.45	51.03	43.27	1.16	37.62	37.33	3.31
RFD-Nets (Our)		**2.4M**	**52.69**	**58.50**	**2.14**	**60.27**	**57.07**	**0.91**	**42.06**	**45.78**	**3.44**
Spline		–	31.82	23.83	4.37	41.05	25.14	2.16	27.91	22.29	6.33
DNN		125.5M	32.03	24.67	4.06	44.23	27.69	2.03	30.31	24.62	5.97
AudioUNet	×4	36.9M	34.12	27.66	9.45	42.50	27.55	9.35	29.28	24.56	8.15
RFD-Nets (Ours)		**2.4M**	**43.10**	**42.57**	**3.71**	**54.26**	**47.09**	**1.86**	**33.25**	**31.16**	**5.72**
Spline		–	28.48	18.29	5.16	37.21	18.77	2.63	25.26	17.88	7.52
DNN		118.0M	30.64	19.64	4.92	40.34	20.64	2.59	26.31	19.65	7.44
AudioUNet	×6	36.9M	33.11	25.98	9.47	42.34	27.29	8.77	28.11	22.62	8.41
RFD-Nets (Ours)		**2.4M**	**38.53**	**34.98**	**4.89**	**49.61**	**39.37**	**2.50**	**30.80**	**27.09**	**7.99**
Method	Scale	Params	Akabera180			SI140			Speaker308		
			PSNR	SNR	LSD	PSNR	SNR	LSD	PSNR	SNR	LSD
Spline		–	38.16	35.26	2.83	50.45	40.51	0.71	42.13	30.23	1.19
DNN		148.1M	40.68	37.09	2.54	52.34	44.80	0.69	43.08	31.95	1.08
AudioUNet	×2	36.9M	42.16	38.95	2.63	53.43	48.47	5.28	44.03	33.11	5.82
RFD-Nets (Ours)		**2.4M**	**47.01**	**49.96**	**2.35**	**62.47**	**60.47**	**0.50**	**56.89**	**54.47**	**0.60**
Spline		–	31.25	23.79	5.43	41.42	25.50	1.41	36.02	19.81	2.07
DNN		125.5M	34.38	26.24	5.36	42.28	27.16	1.35	36.54	20.43	2.00
AudioUNet	×4	36.9M	32.95	26.62	8.72	44.03	29.84	9.50	37.07	21.54	9.69
RFD-Nets (Ours)		**2.4M**	**38.10**	**35.16**	**5.15**	**55.72**	**49.27**	**1.13**	**38.85**	**24.49**	**1.92**
Spline		–	28.35	18.97	6.50	37.50	19.00	1.83	34.32	16.95	2.68
DNN		118.0M	31.66	19.71	6.49	39.78	26.27	1.61	35.17	18.95	2.61
AudioUNet	×6	36.9M	31.99	25.02	8.88	43.94	29.70	8.52	36.32	20.29	8.86
RFD-Nets (Ours)		**2.4M**	**35.57**	**30.95**	**6.95**	**51.29**	**41.91**	**1.57**	**37.50**	**22.26**	**2.56**

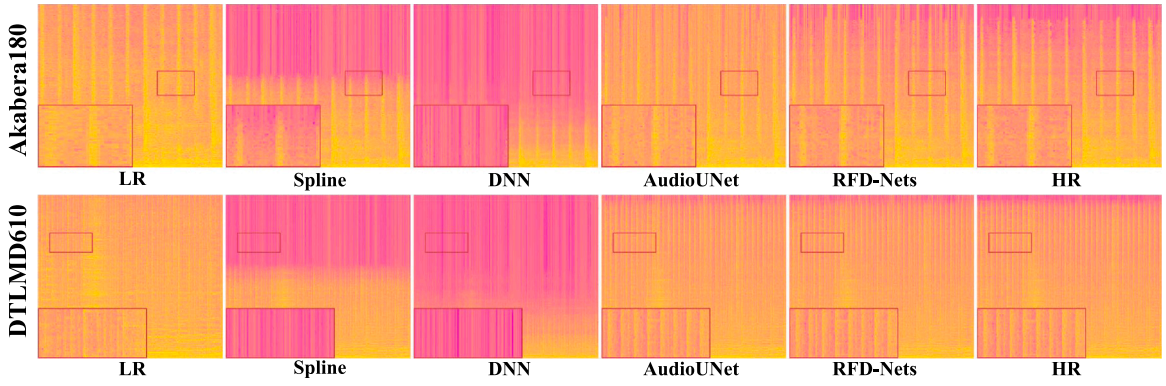

Fig. 4. Reconstruction Spectrogram of models in ×2 scale on Akabera180 and DTLMD610 datasets. LR is the low resolution audio as input for audio super-resolution methods; HR is high resolution audio as groundtruth for audio super-resolution methods; Spline, DNN and AudioUNet are baselines; RFD-Nets is our proposed approach.

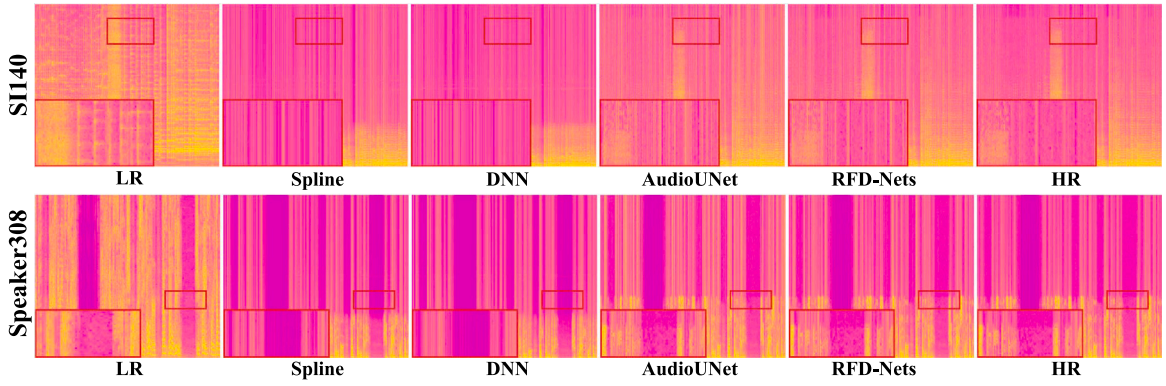

Fig. 5. Reconstruction Spectrogram of models in ×4 scale on SI140 and Speaker308 datasets. LR is the low resolution audio as input for audio super-resolution methods; HR is high resolution audio as groundtruth for audio super-resolution methods; Spline, DNN and AudioUNet are baselines; RFD-Nets is our proposed approach.

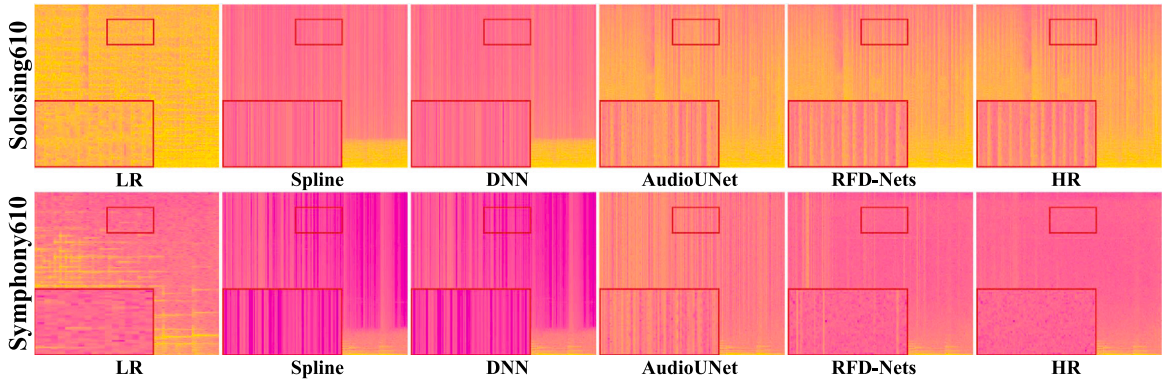

Fig. 6. Reconstruction Spectrogram of models in ×6 scale on Solosing610 and Symphony610 datasets. LR is the low resolution audio as input for audio super-resolution methods; HR is high resolution audio as groundtruth for audio super-resolution methods; Spline, DNN and AudioUNet are baselines; RFD-Nets is our proposed approach.

Table 3
Average PSNR of the bandwidth extension from VCTK and DAPS datasets. The values of PSNR is positively correlated with visual quality, while the value of LSD is negatively correlated with visual quality.

Methods	Input SR	Params	VCTK dataset		DAPS dataset		Akabera180 dataset	
			PSNR	LSD	PSNR	LSD	PSNR	LSD
LP	8k	–	15.74	4.06	15.78	5.00	12.34	9.58
Spec		9.5M	26.19	2.42	36.26	3.06	31.33	7.92
Time		13.6M	22.99	2.03	31.60	2.82	30.16	8.37
WaveNet		6.3M	32.16	2.27	30.14	2.99	31.23	7.93
FFTNet		3.2M	36.33	2.00	35.38	2.80	34.28	7.21
HiFi-GAN+		8.9M	33.53	2.13	30.60	2.80	32.35	7.77
TFiLM		1.5M	35.26	2.09	35.18	2.97	34.06	7.51
RFD-Nets-S (Ours)		**1.5M**	**35.94**	**2.02**	**36.07**	**2.66**	**34.68**	**7.22**
RFD-Nets-B (Ours)		**2.4M**	**36.81**	**1.98**	**36.39**	**2.52**	**35.57**	**6.95**
LP	16k	–	15.74	3.83	13.73	4.61	12.57	9.49
Spec		9.5M	35.74	2.06	40.65	2.58	34.96	7.26
Time		13.6M	29.90	1.92	31.07	3.10	30.11	8.43
WaveNet		6.3M	31.42	1.89	29.05	2.63	31.78	8.06
FFTNet		3.2M	40.59	1.67	39.62	2.44	39.58	4.66
HiFi-GAN+		8.9M	32.16	1.83	29.28	2.35	33.82	7.53
TFiLM		1.5M	39.85	1.71	38.46	2.59	38.97	5.38
RFD-Nets-S (Ours)		**1.5M**	**40.93**	**1.67**	**39.89**	**2.36**	**41.08**	**4.47**
RFD-Nets-B (Ours)		**2.4M**	**42.06**	**1.59**	**41.74**	**2.25**	**42.34**	**3.61**

Comparison with the state-of-the-art methods To demonstrate that the proposed RFD-Nets has stronger robustness and generalization performance, we also conduct a comparative evaluation with the state-of-the-art audio super-resolution algorithms, including the LP (Bachhav et al., 2018), Spec (Eskimez and Koishida, 2019), Time (Li et al., 2019a), FFTNet (Feng et al., 2019), WaveNet (Shen et al., 2018), TFiLM (Birnbaum et al., 2019), and HiFi-GAN+ (Su et al., 2021). To be fair in comparison, we produce their results using public-available implementations provided by the reported settings (Su et al., 2021). In this experiments, we evaluate the compared methods from both 8 kHz to 48 kHz and 16 kHz to 48 kHz extensions on the VCTK (Veaux et al., 2017), DAPS (Mysore, 2014), as well as the most challenging Akabera180 datasets.

Table 3 shows the average PSNR and LSD on the VCTK, the DAPS, and Akabera180 test datasets. The proposed RFD-Nets-B (*i.e.*, RFD-Nets-Base, C = 64, C is the number of feature channels) achieves the best performances on both VCTK, DAPS, and Akabera180 datasets. Particularly, compared to FFTNet (Feng et al., 2019), RFD-Nets improve the average PSNR by 0.48 dB and 1.47 dB on the VCTK dataset with input audio resolutions of 8 kHz and 16 kHz, respectively. Also, compared to Spec (Eskimez and Koishida, 2019), RFD-Nets promote the average PSNR by 1.01 dB and 1.09 dB on the DAPS dataset with input audio resolutions of 8 kHz and 16 kHz, respectively. Compared with TFiLM with same model size (1.5 M parameters), the average PSNR of RFD-Nets-S (*i.e.*, RFD-Nets-Small, C = 32) on the six benchmark datasets not only improves by about **1.14 dB**, but also reduces the average LSD by about **0.31**. These comparisons demonstrate the proposed RFD-Nets have stronger robustness and generalization performance.

Comparison of Mean Opinion Scores (MOS) MOS is a way of subjectively evaluating speech quality. This not only requires the accurate expression of the voice signal content, but also makes the audience feel that the voice is clear and clean. To this end, a total of 100 subjects were asked to rate the overall quality of the sounds produced by our model as well as other comparable methods. In a series of MOS tests, subjects were asked to rate the sound quality of the audio produced on a scale of 1 to 5, where 1 = *poor* and 5 = *excellent*. We conducted four validation tests, including the audio super-resolution results of MOS on Akabera180 dataset with scale = 2, on Speaker308 dataset with scale = 4, and on Solosing610 dataset with scale = 6. We use the VCTK dataset to evaluate the results of MOS for both 8 kHz to 48 kHz and 16 kHz to 48 kHz extensions. Moreover, 50 audio segments are randomly selected on each of the above datasets for evaluation. Each subject could evaluate up to 10 and 25 stimuli for the mentioned audio segments, respectively. Testing stimuli were chosen at random and presented to each subject.

The MOS scores are shown in Fig. 7. Fig. 7(a) shows the audio super-resolution task of subjective evaluation MOS score obtained by Spline, DNN, and AudioUNet on Akabera180 dataset with a scaling factor of 2, in which our RFD-Nets significantly outperforms all the above methods by a large margin. Fig. 7(b) and (c) show the audio super-resolution task of subjective evaluation MOS score on Speaker308 and Solosing610 datasets with a scaling factor of 4 and 6, respectively. Visually, all audio super-resolution methods perform well while our

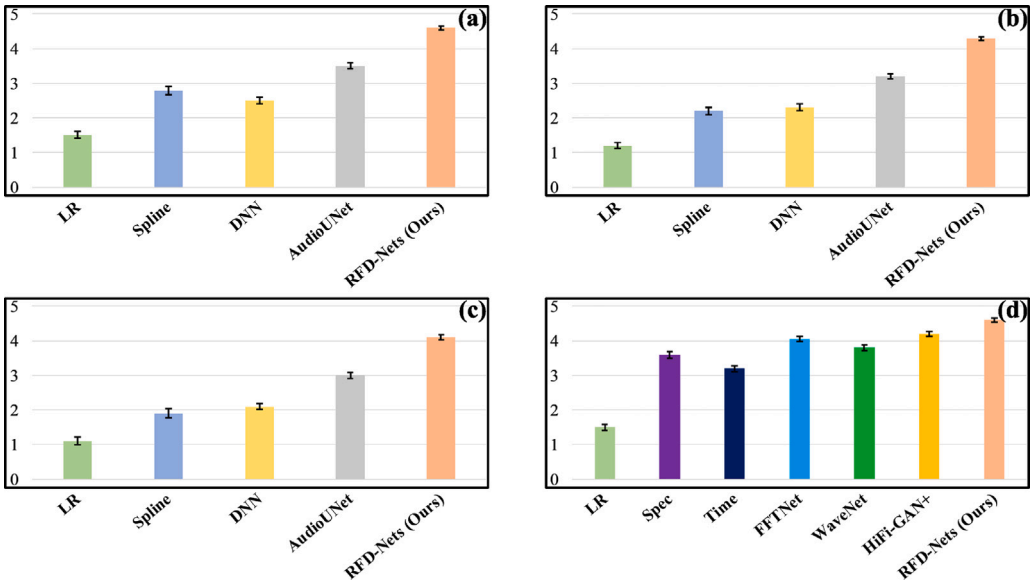

Fig. 7. MOS scores for subjective evaluation of comparable methods. (a) the audio super-resolution results of MOS on Akabera180 dataset with scale = 2; (b) the audio super-resolution results of MOS on Speaker308 dataset with scale = 4; (c) the audio super-resolution results of MOS on Solosing610 dataset with scale = 6; (d) the bandwidth extension results of MOS on VCTK dataset from 16 kHz to 48 kHz.

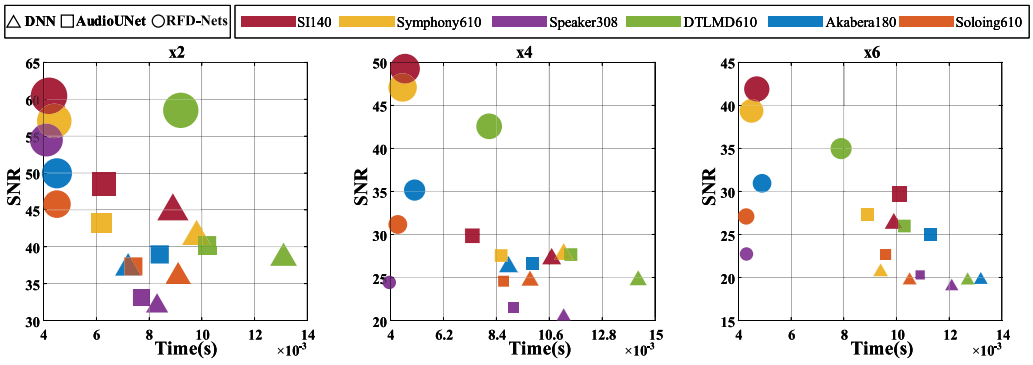

Fig. 8. Comparison of three audio super-resolution network reasoning time. Shapes represent different algorithms, and colors represent different datasets. The size of the shape means the level of SNR value. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

RFD-Nets stands out on Speaker308 and Solosing610 datasets with the highest MOS (4.31) and MOS (4.12), respectively. It is also worth noting that in Fig. 7(d), the proposed RFD-Nets also achieves the best results on the VCTK dataset. Our method achieves state-of-the-art results on both objective evaluation and subjective scores, which demonstrate the effectiveness and superiority of our RFD-Nets.

4.5. Ablation analysis

We quantitatively evaluate the effectiveness of all the components in our RFD-Nets mainly from two aspects: the Feature Diversity (FD) block and the Back-Projection (BP) block. To prove the importance of FD block and BP block in the proposed RFD-Nets, we conduct all degradation experiments on six benchmark datasets with ×4 scale.

In ablation study of the FD block, from Table 4, our FD block increases average PSNR and SNR by about **1.99** dB and **1.12** dB on six benchmark datasets, compared to RFD-Nets without FD block (*i.e.*, Case.c vs Case.b). Moreover, each component in the FD block is vital to improving the performance. *For ablation study of the BP block*, from Table 4, the proposed BP block increases average PSNR and SNR by about **5.93** dB and **6.18** dB on six benchmark datasets, compared to RFD-Nets without BP block. Moreover, the BP block is also vital to improving the performance. It can be observed that lacking any block of RFD-Nets will result in discouraging performance. Particularly, RFD-Nets reach the lowest in most cases in terms of LSD. In addition, we find that the contributions of two blocks to the model are different, FD block shows superior performance on the audio super-resolution task in RFD-Nets. As the feature learning block, FD block increases the diversity of feature learning while ensuring parameter reduction, and provides more detailed knowledge for audio recovery. The effect of the model without FD block is slightly worse, but it is still better than the compared algorithms in Table 2, which indicates the necessity of BP block. Furthermore, the parameters of our model are mainly concentrated in BP block (1.5M/2.4M) for achieving better reconstruction. Generally speaking, both FD block and BP block can facilitate the performance of our framework.

Inference time analysis We compare the inference time of these deep learning models on six datasets in ×2, ×4 and ×6 scales. The results are shown in Fig. 8. Different colors represent different evaluation datasets. Likewise, different shapes represent different compared models. It can be observed that RFD-Nets outperform DNN and AdudioUNet in terms of time and performance. By using fewer parameters, our RFD-Nets display promising performance in the shortest time. Furthermore, to further validate the inference time of RFD-Nets, we report the inference time on dataset VCTK in Table 5. Our RFD-Nets-S (*i.e.*, RFD-Nets-Small, C = 32, C is the number of feature channels) possesses the fastest inference time. In contrast, our RFD-Nets-B (*i.e.*, RFD-Nets-Base, C =

Table 4
Ablation study of different components. PSNR/SNR/LSD values are from six benchmark datasets. The values of PSNR/SNR are positively correlated with visual quality, while the value of LSD is negatively correlated with visual quality. Module detection of our framework in ×4 scale.

Method	Case.	#Params.	DTLMD610			Symphony610			Solosing610		
			PSNR	SNR	LSD	PSNR	SNR	LSD	PSNR	SNR	LSD
W/O FD_BP	a	0.4M	35.16	30.87	7.68	39.75	25.13	5.52	28.64	25.98	8.05
W/O FD_BP[a]	b	2.4M	36.03	31.21	7.54	40.16	26.68	5.09	29.89	26.73	7.72
With FD	c	0.8M	37.29	32.92	6.24	42.65	27.80	4.30	31.51	28.26	7.26
With BP	d	1.5M	40.80	38.75	4.64	50.20	40.34	2.28	32.49	29.90	6.67
RFD-Nets (Our)	e	**2.4M**	**43.10**	**42.57**	**3.71**	**54.26**	**47.09**	**1.86**	**33.25**	**31.16**	**5.72**
Method	Case.	#Params.	Akabera180			SI140			Speaker308		
			PSNR	SNR	LSD	PSNR	SNR	LSD	PSNR	SNR	LSD
W/O FD_BP	a	0.4M	33.32	29.86	8.37	40.23	34.69	4.30	31.51	19.77	5.86
W/O FD_BP[a]	b	2.4M	34.63	30.12	7.96	41.47	35.82	3.71	32.98	20.36	4.14
With FD	c	0.8M	35.08	30.16	7.42	43.64	37.19	3.52	36.94	21.33	3.79
With BP	d	1.5M	37.08	33.46	6.01	51.65	42.50	1.44	38.53	23.92	2.31
RFD-Nets (Our)	e	**2.4M**	**38.10**	**35.16**	**5.15**	**55.72**	**49.27**	**1.13**	**38.85**	**24.49**	**1.92**

[a]Represents the use of convolutional layers to replace FD and BP blocks, and is consistent with the parameters of RFD-Nets.

Table 5
Comparison of the state-of-the-art methods inference time using evaluation dataset VCTK.

Methods	Input SR	Params	VCTK Dataset	
			Time (s)	PSNR
LP	16k	–	–	15.74
Spec		9.5M	0.0191	35.74
WaveNet		6.3M	0.0162	31.42
Time		13.6M	0.0124	29.90
FFTNet		3.2M	0.0103	40.59
HiFi-GAN+		8.9M	0.0089	32.16
TFiLM		1.5M	0.0056	39.85
RFD-Nets-S (Ours)		1.5M	**0.0038**	40.93
RFD-Nets-B (Ours)		2.4M	0.0049	**42.06**

64) has the best balance between the model parameters and inference time. ***This fully illustrates the advantages of recursive operation in RFD in reducing the use of parameters. At the same time, good audio super-resolution reconstruction performance can be achieved, which also shows the effectiveness of FD block in extracting diverse features.***

5. Conclusion

We present a novel lightweight RFD-Nets for accurate audio super-resolution task. We construct the RFD block and the BP block to extract multiple types of features and capture the deep relationships between HR and LR audio pairs for improving the proposed model's performance. Additionally, we present seven different types of complex datasets to comprehensively evaluate the audio super-resolution models, which is crucial for the exploration of the audio super-resolution models in terms of generalization performance. Extensive experiments have shown that our RFD-Nets achieves a commendable balance audio quality and inference time.

CRediT authorship contribution statement

Bo Jiang: Conceptualization, Methodology, Software, Writing – original draft, Visualization, Formal analysis, Validation. **Mixiao Hou:** Writing – review. **Jiahuan Wang:** Writing – review. **Yao Lu:** Writing – review & editing, Supervision. **David Zhang:** Writing – review & editing, Project administration, Funding acquisition. **Guangming Lu:** Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgments

This work was supported in part by the NSFC fund, China (62176077, 61906162), in part by Guangdong Shenzhen joint Youth Fund, China under Grant 2021A151511074, in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2019B1515120055, in part by the Shenzhen Key Technical Project, China under Grant 2020N046, in part by the Shenzhen Fundamental Research Fund, China under Grant JCYJ20210324132210025, in part by the Medical Biometrics Perception and Analysis Engineering Laboratory, Shenzhen, China, in part by Shenzhen Science and Technology Program (RCBS20200714114910193), and in part by Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies, China.

References

Agustsson, E., Timofte, R., 2017. Ntire 2017 challenge on single image super-resolution: Dataset and study. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 126–135.

Bachhav, P., Todisco, M., Evans, N., 2018. Efficient super-wide bandwidth extension using linear prediction based analysis-synthesis. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 5429–5433.

Bansal, D., Raj, B., Smaragdis, P., 2005. Bandwidth expansion of narrowband speech using non-negative matrix factorization. In: INTERSPEECH. Citeseer, pp. 1505–1508.

Birnbaum, S., Kuleshov, V., Enam, S.Z., Koh, P.W., Ermon, S., 2019. Temporal FiLM: Capturing long-range sequence dependencies with feature-wise modulations. In: NeurIPS.

Dong, W., Zhang, L., Shi, G., Wu, X., 2009. Nonlocal back-projection for adaptive image enlargement. In: 2009 16th IEEE International Conference on Image Processing (ICIP). pp. 349–352.

Ekstrand, P., 2002. Bandwidth extension of audio signals by spectral band replication. In: Proceedings of the 1st IEEE Benelux Workshop on Model Based Processing and Coding of Audio (MPCA'02). Citeseer.

Epps, J., Holmes, W.H., 1999. A new technique for wideband enhancement of coded narrowband speech. In: IEEE Workshop on Speech Coding Proceedings. Model, Coders, and Error Criteria. pp. 174–176. http://dx.doi.org/10.1109/SCFT.1999.781522.

Eskimez, S.E., Koishida, K., 2019. Speech super resolution generative adversarial network. In: IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 3717–3721.

Eskimez, S.E., Koishida, K., 2019. Speech super resolution generative adversarial network. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 3717–3721.

Feng, B., Jin, Z., Su, J., Finkelstein, A., 2019. Learning bandwidth expansion using perceptually-motivated loss. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 606–610.

Gray, A., Markel, J., 1976. Distance measures for speech processing. IEEE Trans. Acoust. Speech Signal Process. 24 (5), 380–391. http://dx.doi.org/10.1109/TASSP.1976.1162849.

Guo, Y., Chen, J., Wang, J., Chen, Q., Cao, J., Deng, Z., Xu, Y., Tan, M., 2020. Closed-loop Matters: Dual Regression Networks for Single Image Super-Resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5407–5416.

Haris, M., Widyanto, M.R., Nobuhara, H., 2017. First-order derivative-based super-resolution. Signal, Image Video Process. 11, 1–8.

Irani, M., Peleg, S., 1991. Improving resolution by image registration. CVGIP: Graph. Models Image Process. 53 (3), 231–239.

Jax, P., Vary, P., 2003. Artificial bandwidth extension of speech signals using MMSE estimation based on a hidden Markov model. In: IEEE International Conference on Acoustics, Speech, and Signal Processing. 1, p. I.

Jax, P., Vary, P., 2003. Artificial bandwidth extension of speech signals using MMSE estimation based on a hidden Markov model. In: 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 1, IEEE, p. I.

Jiang, B., Lu, Y., Lu, G., Zhang, D., 2022a. Real noise image adjustment networks for saliency-aware stylistic color retouch. Knowl.-Based Syst. 242, 108317.

Jiang, B., Lu, Y., Wang, J., Lu, G., Zhang, D., 2022b. Deep image denoising with adaptive priors. IEEE Trans. Circuits Syst. Video Technol. 32, 5124–5136.

Jiang, B., Wang, J., Lu, Y., Lu, G., Zhang, D., 2022c. Multi-level noise contrastive network for few-shot image denoising. IEEE Trans. Instrum. Meas..

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Kuleshov, V., Enam, S., Ermon, S., 2017. Audio super resolution using neural networks.

Larsen, E., Aarts, R.M., 2005. Audio Bandwidth Extension: Application of Psychoacoustics, Signal Processing and Loudspeaker Design. John Wiley & Sons.

Lata, K., Dave, M., Nishanth, K.N., 2019. Image-to-image translation using generative adversarial network. In: 2019 3rd International Conference on Electronics, Communication and Aerospace Technology (ICECA). pp. 186–189. http://dx.doi.org/10.1109/ICECA.2019.8822195.

Li, X., Chebiyyam, V., Kirchhoff, K., Amazon, A., 2019a. Speech audio super-resolution for speech recognition. In: INTERSPEECH. pp. 3416–3420.

Li, K., Huang, Z., Xu, Y., Lee, C., 2015. DNN-based speech bandwidth expansion and its application to adding high-frequency missing features for automatic speech recognition of narrowband speech. In: 16th Annual Conference of the International Speech Communication Association. ISCA, pp. 2578–2582.

Li, K., Lee, C.-H., 2015. A deep neural network approach to speech bandwidth expansion. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 4395–4399.

Li, Z., Yang, J., Liu, Z., Yang, X., Jeon, G., Wu, W., 2019b. Feedback network for image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3867–3876.

Lim, T.Y., Yeh, R.A., Xu, Y., Do, M.N., Hasegawa-Johnson, M., 2018. Time-frequency networks for audio super-resolution. In: IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 646–650. http://dx.doi.org/10.1109/ICASSP.2018.8462049.

Mei, Y., Fan, Y., Zhou, Y., Huang, L., Huang, T.S., Shi, H., 2020. Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5690–5699.

Mysore, G.J., 2014. Can we automatically transform speech recorded on common consumer devices in real-world environments into professional production quality speech?—a dataset, insights, and challenges. IEEE Signal Process. Lett. 22 (8), 1006–1010.

Park, K.-Y., Kim, H.S., 2000. Narrowband to wideband conversion of speech using GMM based transformation. In: IEEE International Conference on Acoustics, Speech, and Signal Processing. pp. 1843–1846.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., Devito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A., 2017. Automatic differentiation in pytorch.

Schoenberg, I.J., 1973. Cardinal Spline Interpolation. SIAM.

Seo, H., Kang, H.-G., Soong, F., 2014. A maximum a posterior-based reconstruction approach to speech bandwidth expansion in noise. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 6087–6091.

Shen, J., Pang, R., Weiss, R.J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., et al., 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 4779–4783.

Su, J., Wang, Y., Finkelstein, A., Jin, Z., 2021. Bandwidth extension is all you need. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 696–700.

Tang, H., Liu, H., Sebe, N., 2020. Unified generative adversarial networks for controllable image-to-image translation. IEEE Trans. Image Process. 29, 8916–8929. http://dx.doi.org/10.1109/TIP.2020.3021789.

Veaux, C., Yamagishi, J., MacDonald, K., et al., 2017. Superseded-CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit.

Zhang, K., Gool, L.V., Timofte, R., 2020. Deep unfolding network for image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3217–3226.

Zhao, Y., Wang, R., Jia, W., Wang, W., Gao, W., 2017. Iterative projection reconstruction for fast and efficient image upsampling. Neurocomputing 226, 200–211.

Zou, Z., Lei, S., Shi, T., Shi, Z., Ye, J., 2020. Deep adversarial decomposition: A unified framework for separating superimposed images. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12803–12813. http://dx.doi.org/10.1109/CVPR42600.2020.01282.